Citi Bike Project

Bianca Brusco[1], Michael Sampson[2], and Gokmen Dedemen[1]

[1]New York University (NYU)
[2]NYU Center for Urban Science & Progress

November 8, 2017

Abstract

Bike sharing has gained in popularity as a mean of transportation in urban systems. In New York City, data from Citi Bike usage is publicly available, so trends in riderships can be investigated. In this project, we use one month's data to examine riding trends for Citi Bike subscribers and occasional users of the service, to understand whether there is a difference in the likelihood of taking shorter trips between the two groups. The results show that subscribers are indeed more likely to take shorter trips. One possible explanation of this result is that subscribers are choosing Citi Bike as a mode of transport to cover shorter, so called 'last-mile', segments of their commute more frequently than occasional users. [even when needing to cover shorter distances]

Introduction

New York City introduced a bike-sharing system in 2013: Citi Bike. The service can be used either by subscribing with an yearly membership or by purchasing an occasional pass. The two groups of users are defined as subscribers and as customers. In this project, we investigate whether the ratio of trips longer than average to trips shorter than average is smaller for subscribers than for customers. Indeed, the first group might be more likely to use Citi Bike for shorter trips, as there is no extra charge per trip, while customers might decide to get a pass only to cover more substantial distances.

Data

Citi Bike data is publicly available at https://www.citibikenyc.com/system-data. For this project, we use ridership data from one single month: January 2015. Processing of the data was completed with Pandas for Python 3.

The dataset includes 28,5552 observations, 27,9924 of which are categorized as subscribers and 5,628 as customers.

The observed distribution of trip duration for the two groups is shown in Figures 1 and 2. A brief visual examination of these distribution plots suggest that trip duration for trips made by subscribers and customers may differ in an important way.

Methodology

We compute the mean trip duration for the whole sample, which is 10.96 minutes. Therefore, we define a long trip to be a trip with duration longer than 11 minutes, and a short trip to be a trip with duration

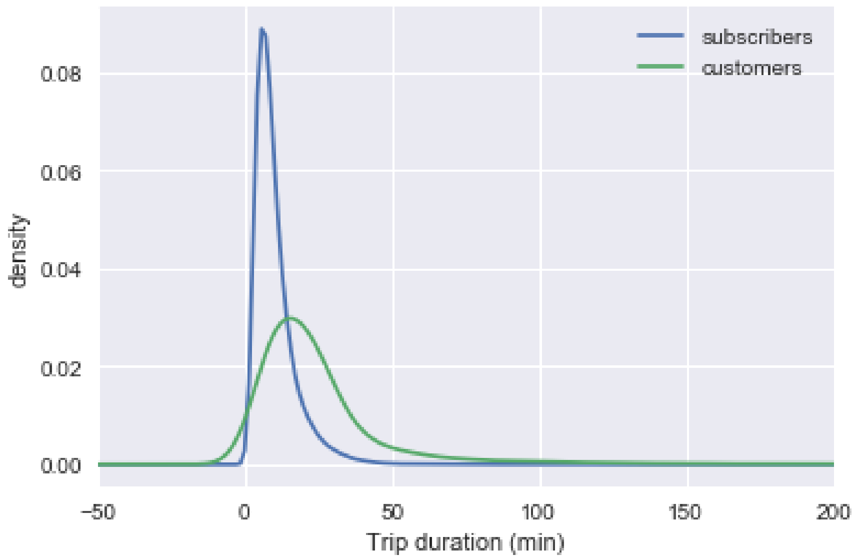

Figure 1: Distribution of trip duration for subscribers and customers.

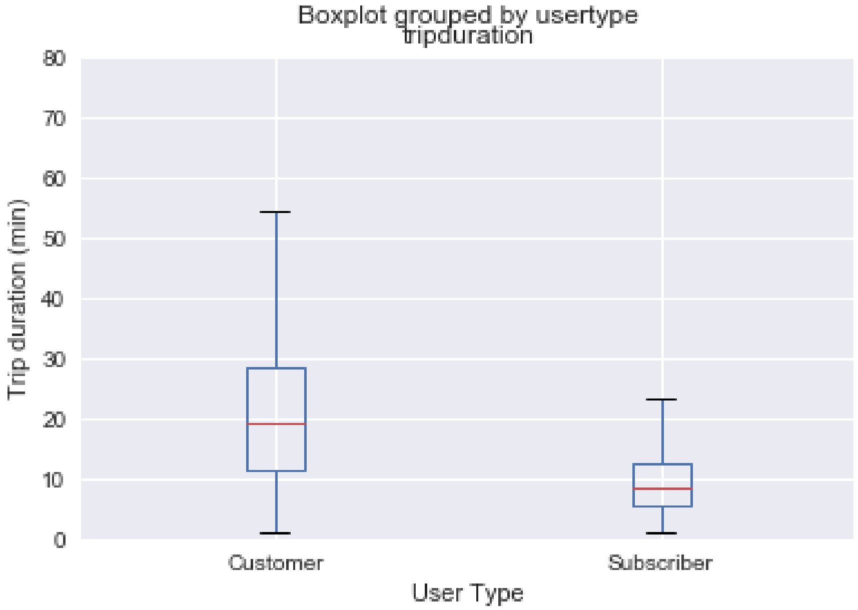

Figure 2: Box plot for trip duration, by user type.

shorter than 11 minutes.

We test the null hypothesis that the mean trip duration for subscribers is equal or longer than the mean trip duration for consumers, against the alternative hypothesis that the mean trip duration for subscribers is shorter than for customers. We use a significance level of $\alpha = 0.05$

Indeed, we are testing

$H_0 : \frac{S_{long}}{S_{short}} \geq \frac{C_{long}}{C_{short}}$

$H_A : \frac{S_{long}}{S_{short}} < \frac{C_{long}}{C_{short}}$

Where S_{long} is the number of long trips taken by subscribers and S_{short} is the number of short trips taken by subscribers. Similar notation holds for customers.

To investigate the difference in means, we use a Z-test, as we are investigating whether two sample proportions appear to come from the same population or not.

An alternative would be to use a Chi-square test. But since we are only testing two proportions, the results should be the same for both tests.

Conclusions

The samples we have used from the population have different proportions of longer to shorter trips.

We can observe the distribution of customer and subscriber trips by trip duration (long and short trips) with the bar plots below, first for absolute counts (Fig. 3) and then normalized (Fig. 4)

We can also observe the distribution of shorter and longer trips by user group with the bar plots below, first for absolute counts (Fig. 5) and then normalized (Fig. 6).

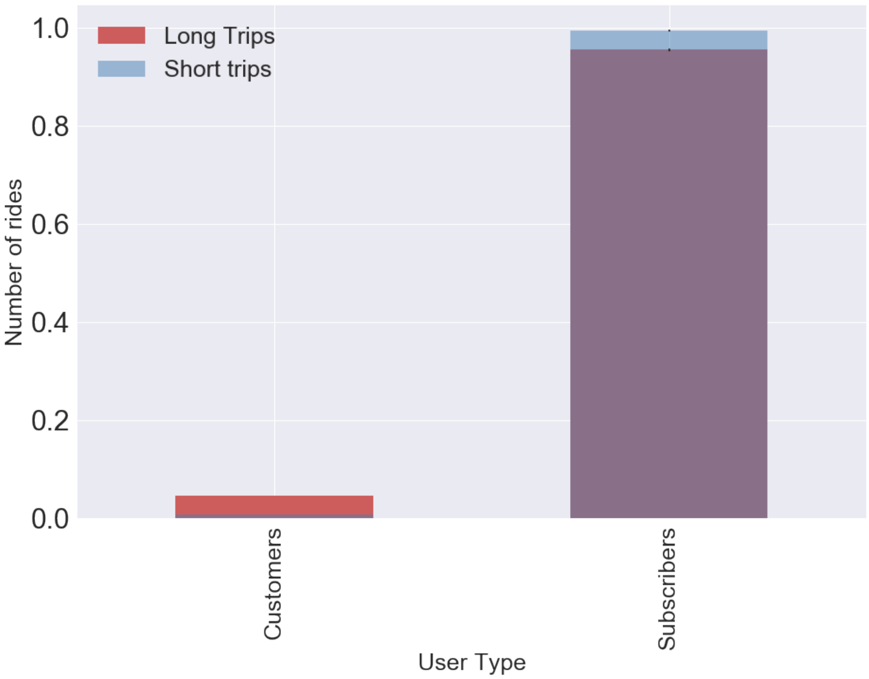

Figure 3: Distribution of Customer and Subscriber Trips by Trip Duration, absolute counts with statistical errors

We see that while subscribers take more shorter trips, customers take more longer trips.

We see again that while subscribers take more shorter trips, customers take more longer trips.

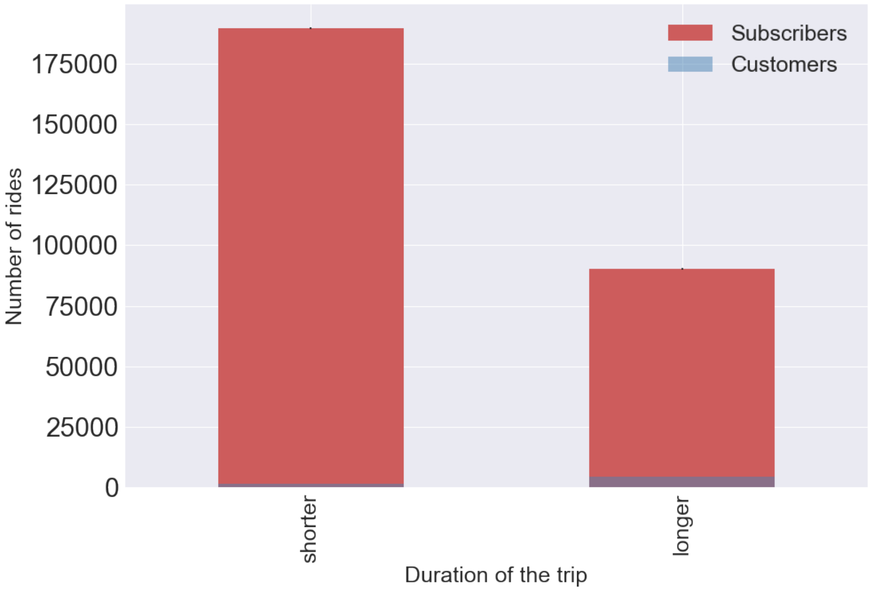

Figure 4: Distribution of Citibike bikers by user type in January 2015, absolute counts, with statistical errors

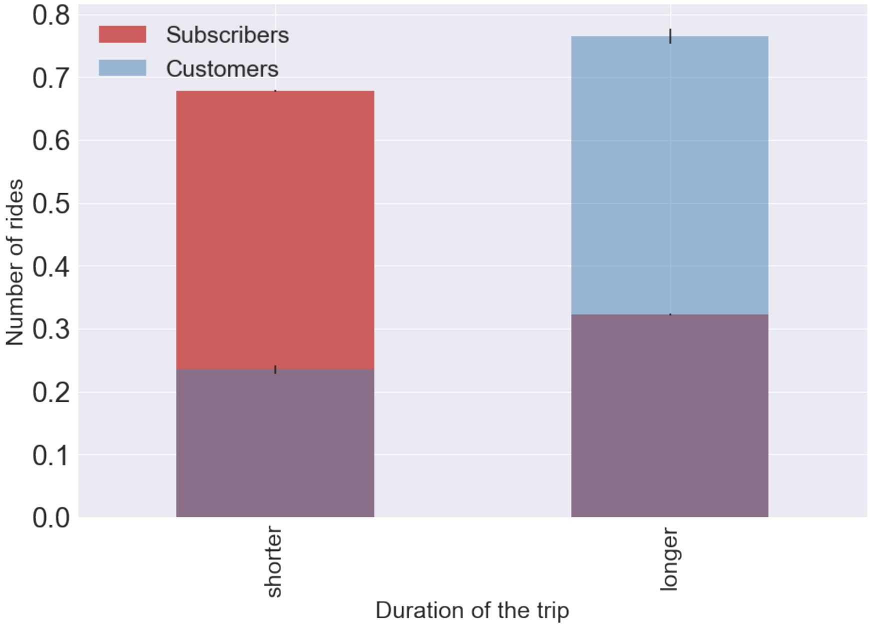

Figure 5: Distribution of Citibike bikers by user type in January 2015, normalized.

We conduct a Z test comparing the two proportions, and obtain a Z-score with absolute value of 414.45. This Z-score is significantly larger than the Z-score for 0.05 significance level, which is 2.96. Therefore, we reject our null hypothesis, and conclude that the ratio of longer/shorter trips is smaller for subscribers than for customers.